

Akhil Reddy Edla

Evansville, IN | +1(913) 327 - 0543 | akhilreddyedla99@gmail.com | [LinkedIn](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

Data Engineer with 5+ years of hands-on experience **designing, building, and scaling** modern data platforms across **AWS, Azure, and GCP**. Skilled in architecting **production-grade batch** and **streaming** pipelines using **Spark, PySpark, Kafka, Pub/Sub, and Delta Lake**, and experienced with lakehouse platforms like **Databricks, Snowflake, Redshift, Synapse, and BigQuery**. Strong in **SQL/Python** development, workflow orchestration (**Airflow, ADF, Composer**), and **CI/CD automation** using **Github Actions, Cloud Build, Azure DevOps, and Terraform**. Brings a deep focus on **governance, lineage, reliability, and observability** using tools such as **Atlan** and **Monte Carlo**, while partnering with **analytics and business** teams to deliver curated, high-quality datasets, ML-ready layers, and BI products that enhance operational efficiency and decision-making.

TECHNICAL SKILLS

Data Engineering & ETL: Airflow, ADF, dbt, Talend, AWS Glue, Cloud Data Pipelines, Lakehouse design, Delta Lake

Big Data & Streaming: spark (PySpark, Spark SQL, Structured Streaming), Kafka, Pub/Sub, Flink, Hive, Hadoop

Cloud Platforms:

- **AWS:** S3, Redshift, Glue, EMR, Lambda, EC2, IAM, Kinesis, FireHose.

- **GCP:** BigQuery, Dataflow, Dataproc, Composer, DataStream, Cloud Functions, Cloud Run, Pub/Sub, Vertex AI

- **Azure:** Data Factory, Synapse, Azure SQL, Blob Storage, Databricks (cluster policies, Unity Catalog, MLflow)

Programming & Scripting: Python, SQL, Shell Scripting, Java.

Databases & Warehouses: snowflake, BigQuery, Redshift, Oracle, PostgreSQL, SQL Server, MySQL, dimensional data modeling & schema design.

DevOps & Infra: Docker, Kubernetes, Terraform, Jenkins, Git/Github, CI/CD automation, cloud monitoring & logging.

DataOps & Observability: Atlan, Unity Catalog, Monte Carlo, metadata automation, lineage tracking, data quality frameworks.

Reporting & Visualization: Power BI, Looker, Tableau, Looker Studio.

PROFESSIONAL EXPERIENCE

Data Engineer

Oct 2024 - Present

CenterPoint Energy | Evansville, IN

- Built and optimized ETL pipelines using **Python, PySpark**, and **SQL** to process structured and unstructured datasets, enhancing analytics query speed by 40% for business teams.
- Leveraged GCP **Pub/Sub, Cloud Functions, Cloud Run, and DataStream** to stream millions of events daily, enabling near real-time insights for enterprise dashboards.
- Automated replication of 150 Coupa tables from **Snowflake** to **BigQuery**, ensuring reliable daily updates while cutting manual refresh work by 70%.
- Integrated SAP **SLT** workflows with GCP infrastructure to maintain synchronized enterprise data, improving consistency across downstream reporting systems.
- Developed Adobe Analytics ingestion pipelines into **GCS** and **BigQuery**, accelerating marketing and product reporting with real-time visibility.
- Created **Power BI** dashboards to **monitor pipeline** performance and data quality, strengthening transparency and trust in key datasets.
- Partnered with business teams to optimize workflows and data models, delivering actionable datasets that reduced reporting delays by 30%.
- Implemented **CI/CD pipelines** using **Github**, Cloud Build, and Terraform, streamlining deployments and reducing production errors by 60%.
- Utilized **Atlan** to maintain and enhance **dataset documentation** by adding detailed column descriptions, business context, lineage notes, and tags, improving data clarity for analysts and downstream teams.
- Created and updated **Atlan metadata** for **BigQuery** tables across multiple pipelines, ensuring consistent naming standards, ownership assignments, and improved data discoverability.
- Leveraged **Monte Carlo** to configure and refine data quality monitors for freshness, volume, and schema changes, strengthening reliability and early detection of anomalies across key datasets.
- Collaborated with data analysts to validate dataset definitions, maintain data dictionaries, and ensure **governance standards** were followed across new and existing pipelines.
- Developed data ingestion and preprocessing workflows for an **AI pipeline** designed to **detect potential pipe corrosion** around gas meters using prompt-based LLM techniques and historical dataset patterns.
- **Maintained comprehensive documentation** and **knowledge transfer**, reducing onboarding time for new engineers by 50%.

Data Engineer

Jan 2023 - Oct 2024

Fifth Third Bank | Evansville, IN

Built and led large-scale data engineering initiatives across **Databricks, AWS, and Azure**, while driving analytics, governance, and ML modernization for enterprise teams.

- Migrated critical workloads from Teradata to **Snowflake, redesigning schemas** and optimizing ETL to reduce compute **cost** and **improve performance**.
- Designed large-scale **Databricks** Lakehouse pipelines using **Spark, Delta Lake, and SQL**, cutting ETL runtimes and increasing pipeline reliability.
- Built Kafka to Databricks **Structured Streaming** pipelines for **real-time fraud detection**, reducing fraud-response times significantly.
- Developed POCs to demonstrate Databricks capabilities in ML, analytics, and scalable ETL accelerating enterprise adoption.

- Managed **Unity Catalog** rollout and created cluster policies for Dev/Prod, improving security, cost control, and governance for 800+ users.
- Implemented **ML workflows** using MLflow for experiment tracking and production model versioning.
- Automated ETL testing + validation using Python, PySpark, and CI pipelines, improving data accuracy and reducing manual QA.
- Built **PySpark** analytics pipelines for **opioid-risk detection** and **healthcare-claim fraud**, supporting compliance and risk management.
- Migrated SQL Server workloads into **AWS S3/Redshift**, enabling cross-platform analytics layers.
- Created APIs to expose curated datasets, enabling faster consumption by downstream teams.
- Implemented Azure DevOps + GitHub CI/CD pipelines for Databricks, Snowflake, and Python services, reducing deployment failures.
- Built observability frameworks and alerting, stabilizing pipeline uptime to 99% across batch and streaming workloads.
- Integrated **Kinesis Data Streams** into existing Spark/PySpark streaming pipelines to handle real-time ingestion and replay scenarios, improving fault tolerance during upstream outages.
- Implemented **Firehose transformations and delivery rules** to clean, enrich, and partition incoming data before landing it in S3, reducing post-processing time for ETL pipelines.

Data Engineer

Kapital Info Tech | India

Jul 2020 - Dec 2022

- Debugged and optimized **Spark** workloads by analyzing stage metrics, **fixing skewed joins**, and improving **partition strategies**, significantly **reducing processing time**.
- Collaborated closely with product and analytics teams to translate ambiguous business logic into reliable **SQL** and **PySpark** transformations.
- Strengthened streaming pipelines by adding **checkpointing**, **DLQ(dead letter queue)**, and **recovery logic** so jobs could handle API failures and resume without manual intervention.
- Tuned Kafka ingestion pipelines to reduce consumer lag during traffic spikes, ensuring real-time dashboards stayed accurate.
- Refactored legacy ETL Python scripts into cleaner, modular components, improving maintainability and simplifying onboarding for new team members.
- Added **robust logging**, metrics, and alerting to critical pipelines, improving observability and cutting root-cause analysis time for on-call support.
- Partnered with **DevOps** to containerize ETL workloads and deploy them on GCP, reducing release friction and improving environment consistency.
- Led weekly **data quality validation** sessions with BI teams to reconcile mismatches, **tighten schema** controls, and improve trust in published datasets.
- Increased **Airflow DAG** stability by adding retries, sensors, and proper dependency handling to avoid failures caused by upstream delays.
- Authored internal documentation that **captured edge cases**, API behavior, **reducing repeated troubleshooting** across the team.
- Supported workflow scheduling, schema governance, stakeholder communication, and design **documentation**.

PROJECTS

- **Snowflake to BigQuery Migration:** Automated schema mapping, incremental ingestion, and validation using Cloud Run + Python.
- **SAP HANA to BigQuery Pipeline:** Ingested SAP datasets using SQL extraction + Cloud Composer workflows.
- **SQL Server to BigQuery Replication:** Designed low-latency replication using DataStream + streaming monitors.
- **Adobe Analytics Automation:** Processed TSV feeds into BigQuery using GCS, Python, and Cloud Scheduler.
- **Event-Driven File Ingestion:** Built Eventarc + Cloud Functions pipeline for real-time GCS to BigQuery ingestion with schema inference and data validation.

EDUCATION

Master of Science in Computer Science

Jan 2023 – May 2024

University of Central Missouri | Warrensburg , MO

CERTIFICATIONS

- [GCP Data Engineer Professional Certification](#)
- [AWS Academy Cloud Architecting](#)
- [AWS Academy Graduate - Cloud Foundations](#)
- Scrum/Agile Methodology Certification
- Databricks Fundamentals
- Power BI for Data Analysts